Sources of Hallucination by Large Language Models on Inference Tasks

Year: 2023 | Citations: 236 | Authors: Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson

---

Abstract

Large Language Models (LLMs) are claimed to be capable of Natural Language Inference (NLI), necessary for applied tasks like question answering and summarization. We present a series of behavioral studies on several LLM families (LLaMA, GPT-3.5, and PaLM) which probe their behavior using controlled experiments. We establish two biases originating from pretraining which predict much of their behavior, and show that these are major sources of hallucination in generative LLMs. First, memorization at the level of sentences: we show that, regardless of the premise, models falsely label NLI test samples as entailing when the hypothesis is attested in training data, and that entities are used as ``indices'' to access the memorized data. Second, statistical patterns of usage learned at the level of corpora: we further show a similar effect when the premise predicate is less frequent than that of the hypothesis in the training data, a bias following from previous studies. We demonstrate that LLMs perform significantly worse on NLI test samples which do not conform to these biases than those which do, and we offer these as valuable controls for future LLM evaluation.